10:50:08 OCA PAD AMENDMENT - PROJECT HEADER INFORMATION 01/20/95

Active

Project #: E-24-X63 Cost share #: Rev #: 1
Center # : 10/24-6-R8288-0A0 Center shr #: OCA file #:
Work type : RES
Contract#: 3032L0014-3B Mod #: 1 Document : CONT
Prime #: W-7405-ENG-36 Contract entity: GTRC

Subprojects ? : N CFDA:
Main project #: PE #:

Project unit: ISYE Unit code: 02.010.124
Project director(s):
PEDERSON S P ISYE (404)894-4962

Sponsor/division names: LOS ALAMOS NATIONAL LAB / UNIVERSITY OF CALIFORNIA
Sponsor/division codes: 240 / 005

Award period: 940909 to 950408 (performance) 950408 (reports)

Sponsor amount	New this change	Total to date
Contract value	0.00	15,739.00
Funded	0.00	15,739.00
Cost sharing amount		0.00

Does subcontracting plan apply ?: N

Title: 1994 WORK ON MCNP STATISTICAL ESTIMATORS

PROJECT ADMINISTRATION DATA

OCA contact: Anita D. Rowland 894-4820

Sponsor technical contact

R.A. FORSTER
(505)667-5777

UNIVERSITY OF CALIFORNIA
LOS ALOMOS NATIONAL LABORATORY
P.O. BOX 1663, M/S B226
LOS ALOMOS, NM 87545

Sponsor issuing office

ALISON BAILEY
(505)665-3900

UNIVERSITY OF CALIFORNIA
LOS ALOMOS NATIONAL LABORATORY
P.O. BOX 1663, M/S P274
LOS ALOMOS, NM 87545

Security class (U,C,S,TS) : U ONR resident rep. is ACO (Y/N): N
Defense priority rating : supplemental sheet
Equipment title vests with: Sponsor GIT
N/A
Administrative comments -
MOD 1 AWARDS A 3-MO NCE THRU 4/8/95

GEORGIA INSTITUTE OF TECHNOLOGY
OFFICE OF CONTRACT ADMINISTRATION

NOTICE OF PROJECT CLOSEOUT

SR 154 2

Closeout Notice Date 04/19/95

Project No. E-24-X63 Center No. 10/24-6-R8288-0A0

Project Director PEDERSON S P School/Lab ISYE

Sponsor LOS ALAMOS NATIONAL LAB/UNIVERSITY OF CALIFORNIA

Contract/Grant No. 3032L0014-3B Contract Entity GTRC

Prime Contract No. W-7405-ENG-36

Title 1994 WORK ON MCNP STATISTICAL ESTIMATORS

Effective Completion Date 950408 (Performance) 950408 (Reports)

Closeout Actions Required:	Y/N	Date Submitted
Final Invoice or Copy of Final Invoice	Y	
Final Report of Inventions and/or Subcontracts	Y	
Government Property Inventory & Related Certificate	N	
Classified Material Certificate	N	
Release and Assignment	Y	
Other	N	

Comments

Subproject Under Main Project No.

Continues Project No.

Distribution Required:

Project Director	Y
Administrative Network Representative	Y
GTRI Accounting/Grants and Contracts	Y
Procurement/Supply Services	Y
Research Property Managment	Y
Research Security Services	N
Reports Coordinator (OCA)	Y
GTRC	Y
Project File	Y
Other	N
	N

NOTE: Final Patent Questionnaire sent to PDPI.

E-24-X63

/

Progress Report for Project No. E-24-X63 (for the period 9-07-94 to 9-30-94)

Shane P. Pederson

February 6, 1995

1 Expenses

The expenses for this period were 1/3 of Dr. Pederson's salary for the period September 19, 1994 to September 30, 1994. Upcoming expenses for the period October 1, 1994 to October 31, 1994, are expected to be 1/3 of Dr. Pederson's salary for this period.

2 Progress Description

In this period Dr. Pederson obtained results regarding the use of the variance of variance (VOV) estimator as an indicator of confidence interval validity. In particular, simulation results were obtained that indicated the fourth moment estimator is the single best predictor of coverage rate validity. By using a classification tree approach, it was found that VOV's less than 0.03 correspond to near nominal coverage for the range of problems considered (both theoretical distributions and a simple Monte Carlo neutron transport two-state problem). This rule was found to work even in cases when the underlying fourth moment quantity does not (or is not indicated to) exist; caution must be made before using this further generalization, and future work will focus on determining the consequences of this result.

An RA for the project was not yet found because notification of award was not received until September 26, 1994.

E-24-X63
2

Progress Report for Project No. E-24-X63 (for the period 10-01-94 to 10-31-94)

Shane P. Pederson

February 6, 1995

1 Expenses

The expenses for this period were 1/3 of Dr. Pederson's salary for the period October 1, 1994 to October 31, 1994. Upcoming expenses for the period November 1, 1994 to November 30, 1994, are expected to be 1/3 of Dr. Pederson's salary for this period.

2 Progress Description

Work continued on the analysis of the variance of the variance indicator, as well as that of other indicators of coverage rate validity. Secondary indicators of validity were determined to generally be other functions of either the variability of the sample variance (that part not measured directly by VOV) or measures of the skewness of the problem. A set of recommendations for Monte Carlo practitioners was proposed, based on the results of the simulation of theoretical and actual (two-state problem) underlying distributions. These recommendations are designed for general used, and rely primarily on the use of the third- or fourth-moment estimate (VOV), the slope estimator, and log-log empirical density plots. The recommendations are designed to indicate when it is likely that the sample is large enough that confidence intervals for the underlying mean will have near-nominal coverage rates. Even if "optimal" conditions are satisfied, it is suggested that sampling continue for an additional length of time (half again as long) to ensure that the indicators discussed above are not biased low. Preparation of a manuscript detailing these results neared completion.

Salim Ur-Rehman was selected as the research associate (1/3 time) for the project. As he had support for Fall quarter, it was decided to ask for a no-cost extension of the project and support Mr. Rehman during Winter quarter, 1995.

E-24-X63
3

Progress Report for Project No. E-24-X63 (for the period 11-01-94 to 11-30-94)

Shane P. Pederson

February 6, 1995

1 Expenses

The expenses for this period were 1/3 of Dr. Pederson's salary for the period November 1, 1994 to November 30, 1994. Upcoming expenses for the period December 1, 1994 to December 31, 1994, are expected to be 1/3 of Dr. Pederson's salary for the period December 1, 1994 to December 9, 1994.

2 Progress Description

Work continued on the manuscript detailing the results of the simulation of indicators of coverage rate validity. A "real" Monte Carlo problem was run by Art Forster of group X-6 at LANL, to test the proposed methodology on a particularly difficult sampling situation. Other work on the manuscript consisted of revising and editing text, and preparing graphs and tables.

Work began on the other main topic of the grant, that is, further research into estimators of the slope of the tail of the log density of the random variables in question. Goodness-of-fit estimators based on a chi-squared statistic and on the Kolmogorov-Smirnov statistic were proposed. Furthermore, the problem of how to choose the fraction of data used to estimate the slope was formulated into three cases: (1), using a constant fraction of the data, usually 10%, 5%, or 1%; (2), using a fraction proportional to a power of the sample size n (such as the square root of n); and (3), using a constant number of observations. Estimators to be considered will be the maximum likelihood estimator from a generalized Pareto distribution, a non-parametric mean estimator, two moment-based estimators, and a slope estimator analogous to that used in linear regression. Future work by the graduate research assistant, to be performed in Winter quarter 1995, will assess these estimators with the goal of determining when the power law description of tail behavior is appropriate.

E-24-X63

4

Progress Report for Project No. E-24-X63 (for the period 12-01-94 to 12-31-94)

Shane P. Pederson

February 6, 1995

1 Expenses

The expenses for this period were 1/3 of Dr. Pederson's salary for the period December 1, 1994 to December 9, 1994, and 1/3 of Graduate Research Assistant Salim Ur-Rehman's support for the period of December 1994 that is covered in Winter quarter 1995 budgeting. Upcoming expenses for the period January 1, 1995 to January 31, 1994, are expected to be 1/3 of Mr. Rehman's support for January 1995, and some money used for the purchase of computer manuals for Mr. Rehman.

2 Progress Description

Final work was completed on the manuscript describing the simulation of several problems, and the use of the variance of the variance (VOV) and other statistics to indicate when likely convergence is achieved.

Simulation studies were conducted to understand the basic properties of estimators of tail slope behavior when they are fit to the same model that generates the data. These confirmatory runs are used as baselines for comparing the results of simulations using other distributions. It was found that the maximum likelihood and probability-weighted moment estimator (PWME) performed the best when Pareto random variables are generated and fit; however, the PWME is of use only when the relevant number of moments actually exist.

It was found that using the tridirectional problem suggested by Tom Booth of X-6, LANL, results in a nonparametric estimator that has a distinctly bimodal distribution over a large range of sample sizes, when the underlying tridirectional distribution has nearly infinite variance. This indicates the dependence of slope estimators on both (a) very large observations, and (b) the structure of lattice-like discrete distributions. This discreteness has implications for the coverage rate indicator study discussed above as well. While common lore has it that as long as two moments exists, the t-statistic will eventually have a limiting standard normal distribution, we

have found that if sufficient discreteness exists (in this case, a repeated inverse sawtooth pattern) convergence to normality can be extremely slowed. Research to be conducted by the supported graduate research assistant will continue on this phenomena, in particular in the development of goodness-of-fit statistics to indicate when it exists. It is expected that at the completion of this project, an additional manuscript characterizing the slope estimators' behavior will be completed by Dr. Pederson and Mr. Rehman.

E-24-X63
5

Progress Report for Project No. E-24-X63 (for the period 1-1-95 to 1-31-95)

Shane P. Pederson

February 6, 1995

1 Expenses

The expenses for this period were 1/3 of Salim Ur-Rehman's support for the month of January 1995, and approximately $100.00 for computer books. The period of the contract was extended until April 8, 1995 to allow charging of Mr. Rehman's time. It is expected that upcoming expenses for the period February 1, 1995 to February 28, 1995 will be 1/3 of Mr. Rehman's support during this time, as well as travel of the principal investigator to Los Alamos National Laboratory, Los Alamos, New Mexico, to confer with funders from group X-6.

2 Progress Description

Properties of tail slope estimators were studied during this month, mainly via computer simulation. It was found that the maximum likelihood and nonparametric slope estimators are relatively insensitive to where the data are truncated (within the interval of the two observations for which the truncation point lies between), as long as the underlying distribution is smooth. When discreteness occurs, and the skewness is severe, for small samples the estimators are sensitive to where truncation is chosen. Preliminary results indicate that a truncation point near the right side of the interval gives the least biased results. Further simulations will be conducted during the following month to confirm this result.

The primary measure of goodness-of-fit was chosen to be the chi-squared statistic, with intervals chosen to have approximately equal spacings on the logarithmic scale. This measure is useful in determining overall lack-of-fit, but is not more sensitive to certain patterns. For that reason, an alternative measure, based on an analogy with linear regression, is also used; this measure attempts to detect nonlinearities in the actual tail log-density; if the lack-of-fit appears to be random, this measure will not detect it, but if a pattern exists (such as *xlog(x)*) it will be detected. The graduate research assistant began simulations to quantify the behavior of these measures.

Another topic of study was the behavior of slope estimators under arbitrary conditions, in particular when there is systematic variation around a continuous power law model. Simuations suggest that the maximum likelihood Pareto estimator may be sensitive to variation in the leftmost part of the distribution, but relatively insensitive to variations in the tails. The other estimators appear to be insensitive to variation on all parts of the range. Further analysis of these results will give information on the bias present in these cases. In addition, results from the tridirectional discrete model will be analyzed in this context.

E-24-X63
6

Progress Report for Project No. E-24-X63 (for the period 2-1-95 to 2-28-95)

Shane P. Pederson

March 6, 1995

1 Expenses

The expenses for this period were 1/3 of Salim Ur-Rehman's support for the month of February 1995. It is expected that upcoming expenses for the period March 1, 1995 to March 31, 1995 will be 1/3 of Mr. Rehman's support during Winter Quarter (which ends March 19), as well as travel of the principal investigator to Los Alamos National Laboratory, Los Alamos, New Mexico, to confer with funders from group X-6.

2 Progress Description

Additional work on the use of moment estimators to predict coverage rate behavior was completed this month. This work considered framing convergence criteria in terms of skewness-modified confidence intervals. Analysis of highly discrete distributions found that not only does discreteness make too few intervals cover the parameter of interest (i.e., the intervals are *anti-conservative*, but that discreteness also makes upper intervals *conservative* after convergence of t to normality has been achieved.

Analysis of one of the standard goodness-of-fit tests, the Kolmogorov-Smirnov procedure, was performed for the Pareto distributions. Mr. Rehman developed tables of probability points for this distribution, correcting an erroneous result in the literature. This correction will be written up and submitted to *IEEE Transactions on Reliability.*

A lack-of-fit estimator, analogous to the lack-of-fit to linearity quantity used in regression, was developed. This measure will be tested in March, via simulation, to determine if it is useful in indicating when non-power law tail behavior is present. Data will be generated from the tridirectional discrete distribution of Booth, as well as power-law data with sinusoidal variation added on.

Work in March will finish the simulations on different methods of sample size acquisition: fixed fraction, fixed amount, and amount proportional to a power of sample size n. While theory suggest that having sample size increase in inverse proportion to $n^{-2/5}$, preliminary results show that a wide

varieties of powers of n give roughly similar results, especially when the model generating the data only approximately follows a power law. This simulation, in concert with simulations on the type of estimator (nonparametric, Pareto-based, probability-weighted moment), will determine the best (i.e., most robust) tail slope estimator for general use.

E-24-X63

7

Progress Report for Project No. E-24-X63 (for the period 3-1-95 to 3-31-95)

Shane P. Pederson

April 7, 1995

1 Expenses

The expenses for this period were 1/3 of Salim Ur-Rehman's support for the month of March 1995.

2 Progress Description

Simulations to investigate the properties of the regression lack-of-fit test and sample size determination were conducted in March. It was found that the lack-of-fit procedure gave results similar to standard goodness-of-fit procedures, with the added benefit of visual display with which the user can easily discover systematic misfitting (even if not statistically significant).

Simulations were also conducted on the determination of sample size fraction to use in estimating the tail of the distribution for Pareto and similar data. It was found that fraction powers of sample size n such as $1/2$ (corresponding to square roots) gave the best combination of information obtained and parsimony. In particular, these sample sizes outperformed the standard method in which a constant fraction of the data are used. Writing up of the simulation results for submission to a journal was also begun in March.

E-24-X63

Final Report for Project No. E-24-X63

Shane P. Pederson

April 8, 1995

1 Introduction

This is a summary report of the research completed on Project number E-24-X63, *Work on MCNP Statistical Estimators*, for Los Alamos National Laboratory (LANL), by the above author. Collaborators on this project were R. A. Forster and T. E. Booth, both of group X-6 at LANL. S. Ur-Rehman served as a Graduate Research Assistant for the project. The work examined the behavior of standard statistical estimators used in Monte Carlo particle transport codes, when output distributions are severely skewed due to the use of variance reduction techniques. Additionally, techniques for estimating the tail of output distributions were studied. Partially as a result of this grant, a research paper was completed and another begun, and some sections of this work will be presented at the Interface on Computer Science and Statistics conference this summer.

2 Work on Standard Procedures

We considered the properties of standard statistical estimators, particularly the mean, of parameters from Monte Carlo output distributions when these distributions are extremely right-skewed. This skewness arises from excessive particle weighting due to multiple applications of variance reduction techniques. We found that the fourth moment estimator, sometimes termed the *variance of the variance*, is an effective indicator of coverage rate validity of standard normal-theory based confidence intervals. This was verified both via theoretical Edgeworth expansion of the distribution of the scaled sample mean t and via computer simulation. In addition, it was found that the use of simple tail slope estimators allows the user of the Monte Carlo code to determine the approximate number of underlying moments that exist. This information is important as it determines whether standard normal-theory expansions for t are adequate. Finally, corrected intervals based on Edgeworth expansions were found to approximately halve the error in confidence interval coverage rates.

The sampling distributions used in simulations were selected to mimic actual Monte Carlo output distributions. Most assume that the tail of the

data follow a *power-law* distribution. A variety of continuous distributions based on the Pareto or Cauchy random variables were used, as well as a flexible discrete-state distribution due to T. E. Booth. It was found that confidence intervals perform most poorly when both (a) low numbers of moments exist, and (b) the underlying sampled distribution is highly discrete in nature.

These results were incorporated into a paper submitted to the *Journal of Statistical Computation and Simulation.* Furthermore, rules based on the above results will be included in the next version of the Monte Carlo particle transport code MCNP™ [1], with corresponding documentation.

3 Properties of Tail Slope Estimators

The second main aspect of this work was to characterize the statistical properties of the tail slope estimators described above. In particular it was found that the standard nonparametric slope estimator is more robust than the estimator arising from the Generalized Pareto distribution, in particular when data are generated from the Booth (highly discrete) distribution. In large samples from smooth distributions these two estimators behave similarly, but in small samples or those from more irregular distributions, the nonparametric estimators were found to be superior. The author recommends that this estimator be used in future work.

A useful companion to any tail slope estimator is a measure of the goodness-of-fit of the power-law model to the tail of the data. Work was begun on this question, with both standard goodness-of-fit (Chi-square and Kolmogorov-Smirnov) tests and a new procedure, based on the analog of a linear regression slope estimator, examined via simulation. Preliminary results indicate that the regression estimator gives answers qualitatively similar to the standard methods. Simulations will continue on this question.

Another important question when fitting a model to the tail of a distribution is the amount of data to be used. Simulations were conducted which examined the strategies of using (a) a constant fraction of the data; (b) a fraction proportional to a power of the sample size (e.g., square root); and (c) a constant *amount* of data. A subset of the distributions described in Section 2 were used in these simulations. In addition, perturbations were used to simulate data from a distribution for which the power law tail model does not hold. Again, results are preliminary, but suggest that method (b) gives the best combination of data parsimony and tail information.

These results are currently being written in a research paper to be submitted to *Technometrics.* When definite conclusions are obtained regarding the goodness-of-fit test, it will be incorporated into a new version of of MCNP. Any code that pertains to these statistical techniques will also be supplied by the author to group X-6 of LANL.

4 Conclusions

We were able to answer the main questions posed at the beginning of the study. We have obtained measures which quantify the coverage rate behavior of confidence intervals for Monte Carlo output distributions, even for highly skewed situations. Heuristic rules have been developed for the generic MCNP user, to aid in determining whether the code has been run sufficiently long for valid statistical inferences to be made. The tail slope estimator has been established as an essential tool in this process. While some questions about its properties remain, these will be answered in the near future. In conclusion, it is hoped that this work will foster additional interactions between researchers and practioners in these fields of data analysis and Monte Carlo particle simulation.

References

[1] J. F. BRIEMEISTER, Ed., "MCNP^{TM} — A General Monte Carlo N-Particle Transport Code: Version 4A," LA-12625-M, Los Alamos National Laboratory (1994).